Biometrical Journal () , zzz–zzz / DOI: 10.1002/bimj.200100000

An empirical Bayes approach to network recovery using external knowledge

Gino B. Kpogbezan[*,1], **Aad W. van der Vaart** [1], **Wessel N. van Wieringen** [2,3], **Gwenaël G. R. Leday** [4] , and **Mark A. van de Wiel** [2,3]

[1] Mathematical Institute, University of Leiden, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
[2] Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
[3] Department of Epidemiology and Biostatistics, VU University Medical Center, PO Box 7057, 1007 MB Amsterdam, The Netherlands
[4] MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Cambridge CB2 0SR, United Kingdom

Received zzz, revised zzz, accepted zzz

Reconstruction of a high-dimensional network may benefit substantially from the inclusion of prior knowledge on the network topology. In the case of gene interaction networks such knowledge may come for instance from pathway repositories like KEGG, or be inferred from data of a pilot study. The Bayesian framework provides a natural means of including such prior knowledge. Based on a Bayesian Simultaneous Equation Model, we develop an appealing Empirical Bayes (EB) procedure which automatically assesses the agreement of the used prior knowledge with the data at hand. We use variational Bayes method for posterior densities approximation and compare its accuracy with that of Gibbs sampling strategy. Our method is computationally fast, and can outperform known competitors. In a simulation study we show that accurate prior data can greatly improve the reconstruction of the network, but need not harm the reconstruction if wrong. We demonstrate the benefits of the method in an analysis of gene expression data from GEO. In particular, the edges of the recovered network have superior reproducibility (compared to that of competitors) over resampled versions of the data.

Key words: Empirical Bayes; High-dimensional Bayesian inference; Prior information; Undirected network; Variational approximation.

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1022/bimj.XXXXXXX (please delete if not applicable)

1. Introduction Many areas of the quantitative sciences have witnessed a data deluge in recent years. This is due to an increased capacity of measuring and storing data in combination with a reduction in costs of acquiring this data. For instance, in the medical field high-throughput platforms yield measurements of many molecular aspects (e.g. gene expression) of the cell. As many as $20,000$ genes of a single patient can be characterized simultaneously. However, although the costs of such techniques have gone down over the years, the number of patients n in a typical clinical study is still small compared to the number of variables p measured. Reliable analysis of data of such a "$n \ll p$" study is difficult. In this paper we try to solve the problem of few replicate measurements by incorporating external (or "prior") data in the analysis. To allow interpretation, we restrict ourselves to predefined subsets of genes (e.g. pathways) for which p is usually moderately larger than n.

High-dimensional modelling based on a small data set is particularly challenging in studies of relationships between variables. The number of potential pairwise relationships between even a modest number of

*Corresponding author: e-mail: g.b.kpogbezan@math.leidenuniv.nl

genes is $p(p-1)/2$. However, some of these relationships may be known from the vast body of medical literature available. For instance, the current beliefs on interactions among genes is condensed in repositories like KEGG and Reactome. Although such information may not be reliable, or be only partially relevant for the case at hand, its flexible inclusion may help the analysis of high-dimensional data. Methodology that exploits such prior information may accelerate our understanding of complex systems like the cell.

The cohesion of variables constituting a complex system is often represented by a network, also referred to as a *graph*. A graph $\mathcal{G}$ consists of a pair $(\mathcal{I}, \mathcal{E})$ where $\mathcal{I} = \{1, ..., p\}$ is a set of indices representing nodes (the variables of the system) and $\mathcal{E}$ is the set of edges (relations between the variables) in $\mathcal{I} \times \mathcal{I}$. An edge can be characterized in many ways, we concentrate on it representing conditional independence between the node pair it connects. More formally, a pair $(i_1, i_2) \in \mathcal{E}$ if and only if random variables represented by nodes i_1 and i_2 are conditionally dependent, given all remaining nodes in $\mathcal{I}$. All pairs of nodes of $\mathcal{I}$ not in $\mathcal{E}$ are conditionally independent given the remaining nodes. Graphs endowed with this operationalization of the edges are referred to as conditional independence graphs (Whittaker, 1990).

Conditional independence graphs are learned from data by graphical models. Graphical models specify how data are generated obeying the relations among the variables as specified by a conditional independence graph. A Gaussian Graphical Model (GGM) assumes data are drawn from a multivariate normal distribution:

$$Y^j \sim^{\text{iid}} \text{N}(0, \Omega_p^{-1}), \qquad j \in \{1, ..., n\}. \tag{1}$$

Here Y^j is a p-dimensional random vector comprising the p random variables $Y_1^j, \ldots, Y_p^j$ corresponding to the nodes of $\mathcal{I}$ and Ω_p^{-1} is a non-singular $(p \times p)$-dimensional covariance matrix. The matrix Ω_p, as opposed to its inverse, is referred to as the *precision matrix*. For a GGM the edge set $\mathcal{E}$ of the underlying conditional independence graph corresponds to the nonzero elements of Ω_p (Lauritzen, 1996). Hence, to reconstruct the conditional independence graph it suffices to determine the non-zeros elements of this matrix.

Reconstruction of the conditional independence graph may concentrate on the direct estimation of the precision matrix. Here we choose a different estimation strategy. This exploits an equivalence between Gaussian graphical models and Simultaneous Equations Models (SEMs), which we introduce first before pointing out the equivalence. Our choice for SEM is mainly motivated by its flexibility and its performance. It can account for experimental or biological covariates in the regression, and extensions to non-Gaussian data are available (Chen et al., 2015; Allen and Liu, 2013; Yang et al., 2012; Ravikumar et al., 2010). Its Bayesian counterpart is appealing for including prior knowledge, which likely is more complicated in many other frameworks. Its good performance in comparison with alternatives including (sparse) graphical models was demonstrated by Leday et al. (2015). In addition, SEM is also computational efficient (Meinshausen and Bühlmann, 2006). We treat SEMs as a system of regression equations, with each equation modelling the conditional distribution of a node given the other nodes. If we collect all observations on node $i \in \mathcal{I}$ in a vector $Y_i := (Y_i^1, \ldots, Y_i^n)^T$, then we can write:

$$Y_i = X_i \beta_i + \epsilon_i, \quad i \in \mathcal{I}, \tag{2}$$

where X_i is the $n \times (p-1)$-matrix with columns the observations of the $p-1$ nodes different from i, i.e. $X_i = [Y_1, Y_2, ..., Y_{i-1}, Y_{i+1}, ..., Y_p]$ (where the square brackets mean "combine the vectors in a matrix"). The error vector ϵ_i is defined by the equation, and possesses a multivariate Gaussian distribution $\text{N}(0, \sigma_i^2 \mathbf{I_n})$ under the GGM. (The covariances between the errors of different equations are in general non-zero, but are left unspecified.) The equivalence between the thus formulated SEM and the GGM as specified above stems from the one-to-one relationship between the regression parameters of the SEM and the elements of the GGM's precision matrix (Lauritzen (1996)): $\beta_{i,r} = -\omega_{ii}^{-1} \omega_{ir}$. In particular, (non)zero entries in the i-th row vector of the precision matrix Ω_p correspond to the (non)zero coefficients of β_i. The problem of identifying (non)zero entries in Ω_p can therefore be cast as a variable selection problem in the p regression models (2). Lasso regression (Tibshirani, 1996) may be used for this purpose (as in Meinshausen and

Bühlmann (2006)), but other variable selection methods have also been employed. The problem that every partial correlation appears in two regression equations is usually resolved by post-symmetrization through application of the 'AND'-rule: an edge $(i,j) \in \mathcal{E}$ if and only if $\beta_{i,j} \neq 0$ and $\beta_{j,i} \neq 0$ (Meinshausen and Bühlmann, 2006). Graph structures recovery based on model (2) performs well and is widely used in practice.

Previously, we proposed a Bayesian formulation of the SEM (Leday et al., 2015). In this Bayesian SEM (henceforth BSEM) the structural model (2) is endowed with the following prior:

$$\begin{aligned} \epsilon_i |\, \sigma_i^2, \tau_i^2 &\sim \mathrm{N}(0_n, \sigma_i^2 \mathbf{I}_n), \\ \beta_i |\, \sigma_i^2, \tau_i^2 &\sim \mathrm{N}(0_s, \sigma_i^2 \tau_i^{-2} \mathbf{I}_s), \\ \tau_i^2 &\sim \mathrm{Gamma}(a_1, b_1), \\ \sigma_i^{-2} &\sim \mathrm{Gamma}(a_2, b_2), \end{aligned} \tag{3}$$

where $\mathbf{I}$ is an identity matrix, $s = p - 1$, and Gamma(a, b) denotes a gamma distribution with shape parameter a and rate parameter b, and τ_i^2 and σ_i^{-2} are independent. The normal-gamma-gamma (NGG) prior of model (3) regularizes the parameter estimates (e.g. estimated as the posterior mean) in two distinct ways. First, due to the normal prior on the regression coefficients $\beta_{i,r}$ (corresponding to a ridge penalty), the estimates of these parameters are shrunken *locally* (i.e. within each equation) to zero. Second, the estimates are simultaneously shrunken *globally* (i.e. across equations), due to the fact that the hyperparameters $\alpha = \{a_1, b_1, a_2, b_2\}$ do not depend on the index i. There seems to be no reason to connect the error variances (which reflect the noise levels of the genes) across the equations, and hence we use a vague prior (e.g. $a_2 = b_2 = 0.001$). In contrast, estimating the parameters a_1, b_1 in EB fashion is advantageous, as it further "borrows information" across the regression equations. The resulting global shrinkage improves inference in particular for large networks (see also Section 5). Note that assuming a Gaussian distribution for the regression coefficients is also done in ridge regression and random effects models. The BSEM model can be fit computationally efficiently by a variational method, and generally outperforms the aforementioned lasso regression approach to the estimation of model (2). Furthermore, variables can be accurately selected based on the marginal posterior distributions of the regression coefficients (Leday et al., 2015).

The problem of network reconstruction is challenging due to the vast space of possible graphs for even a moderate number of variables. This endeavour is further complicated by the inherent noise in the measurements used for the reconstruction. Fortunately, network reconstruction need not start from scratch, as often similar networks have been studied previously. Prior information on the network may be available in the literature, repositories, or simply as pilot data. It is natural to take such information along in network reconstruction. Many works have already been devoted to incorporating prior knowledge into network reconstruction. Among these studies, Imoto et al. (2003) use energy functions to incorporate prior knowledge sources into Bayesian gene regulatory network models and propose the incorporation of many types of different prior knowledge, including literature-based knowledge. The approach of Imoto et al. has been extended by Werhli and Husmeier which proposed a framework to incorporate multiple sources of prior knowledge into dynamic Bayesian network using MCMC sampling (Werhli and Husmeier, 2007). In the same line, Steele et al. proposed an advanced text-mining technique to incorporate literature-based prior knowledge into Bayesian network learning of gene networks. Similarly, Li et al. developed an approach that combines literature mining and microarray analysis in constructing biological networks (Li et al., 2006). Murkherjee and Speed (2008) proposed a method to incorporate network features including edges, classes of edges, degree distributions, and sparsity using MCMC sampling in Bayesian network learning. Still in Bayesian network learning, Isci et al. (2013) proposed also a framework to incorporate multiple sources of external knowledge where the incorporation of external knowledge uses Bayesian network infrastructure itself. However, none of these proposed methods explicitly estimate the agreement of the prior knowledge with the data at hand.

In this paper we develop a method for incorporating external data or prior information into the reconstruction of a conditional independence network. To this aim we extend in Section 2 the Bayesian SEM

framework (2)-(3). The extension incorporates prior knowledge in a flexible manner. Next in Section 3 we develop a variational Bayes approach to approximate the posterior distributions of the regression parameters for given hyperparameters, and show this to be comparable in accuracy to Gibbs sampling, although computationally much more efficient. In Section 4 this is complemented by a derivation of an empirical Bayes approach to estimate the hyperparameters. Using simulations we show in Section 5 that the method performs better, in terms of ROC curves, than BSEM when the prior knowledge agrees with the data, and is as accurate when it is not. In Section 6 we show the full potential of our approach on real data. We conclude the paper with a discussion.

2. Model The BSEM approach, comprising model (2) with priors (3), is modified to incorporate external information on the to-be-reconstructed network. The resulting model is referred to as BSEMed (BSEM with *e*xternal *d*ata).

Prior knowledge on the network is assumed to be available as a "prior network", which specifies which edges (conditional independencies) are present and absent. This is coded in an adjacency matrix P, which contains only zeros and ones corresponding to the absence and presence of an edge in the prior network. That is, $\mathrm{P}_{i,r} = 1$ if node i is connected with node r and $\mathrm{P}_{i,r} = 0$ otherwise. Note that the adjacency matrix P is symmetric (for the purpose of undirected network reconstruction).

The BSEMed approach keeps equation (2), but replaces the priors (3) of BSEM by:

$$\begin{aligned} \epsilon_i|\,\sigma_i^2, \tau_{i,0}^2, \tau_{i,1}^2 &\sim \mathrm{N}(0_n, \sigma_i^2 \mathbf{I}_n), \\ \beta_i|\,\sigma_i^2, \tau_{i,0}^2, \tau_{i,1}^2 &\sim \mathrm{N}(0_s, \sigma_i^2 \mathbf{D}_{\tau_i^{-2}}), \\ \mathbf{D}_{\tau_i^{-2}} &= \mathrm{diag}(\tau_{i,1}^{-2}, ..., \tau_{i,s}^{-2}), \\ \tau_{i,r}^2 = \begin{cases} \tau_{i,0}^2 \sim \mathrm{Gamma}(a_0, b_0), & \text{if} \quad \mathrm{P}_{i,r} = 0, \\ \tau_{i,1}^2 \sim \mathrm{Gamma}(a_1, b_1), & \text{if} \quad \mathrm{P}_{i,r} = 1, \end{cases} & \\ \sigma_i^{-2} \sim \mathrm{Gamma}(a_2, b_2). & \end{aligned} \tag{4}$$

where $\beta_i = \beta_{i,1}, ..., \beta_{i,i-1}, \beta_{i,i+1}, ..., \beta_{i,p}$.

The normal-gamma-gamma-gamma (NGGG) prior (4) retains the ridge-type regularization of the regression parameters $\beta_{i,r}$ of (3), through Gaussian priors on these coefficients. The crucial difference between the two priors reveals itself in the variances of the latter priors. For each regression equation i there are two possible variances:

$$\beta_{i,r} \sim \begin{cases} \mathrm{N}(0, \sigma_i^2 \tau_{i,0}^{-2}), & \text{if} \quad \mathrm{P}_{i,r} = 0, \\ \mathrm{N}(0, \sigma_i^2 \tau_{i,1}^{-2}), & \text{if} \quad \mathrm{P}_{i,r} = 1. \end{cases}$$

Hence, the regression coefficients corresponding to edges that are present according to the prior information share the same variance, and similarly for the other set of regression coefficients. Both variances can be both small and large, as they are themselves modelled through Gamma priors, where small values lead to small regression coefficients. If the prior information on the network were correct, then naturally a small value of $\tau_{i,0}^{-2}$ would be desirable, smaller than the value of $\tau_{i,1}^{-2}$. However, the construction is flexible in that the two values, and even their priors, are not fixed a-priori. In (4) the two parameters $\tau_{i,0}^{-2}$ and $\tau_{i,1}^{-2}$ are assumed to have gamma priors, with different hyperparameters (a_0, b_0) and (a_1, b_1). For further flexibility these hyperparameters will be estimated from the data with an empirical Bayes method. Then, if the absence of an edge in the prior network is supported by the current data, the corresponding regression coefficient $\beta_{i,r}$ may stem from a prior with a small variance, and will tend to be small; a similar, but opposite, situation will occur for edges that are present in the prior network. Indeed in Section 5 we shall see that the EB approach will tend to find similar values of $\tau_{i,0}^2$ and $\tau_{i,1}^2$ when the prior knowledge is non-informative, and rather different values otherwise.

The fact that model (4) contains the model (3) as a submodel, provides robustness against the misspecification of the prior information. Although the number of latent variables in (4) is considerably higher

(namely $p-1$ additional variances, one for each regression equation), the actual number of extra parameters is only two (the pair (a_1, b_1)). This suggests that if the prior information doesn't agree with the data at hand, then the cost in terms of precision of the estimators is minor. It is amply compensated by the gains if the prior information is correct. We corroborate this in our simulation study in Section 5. In this connection it is also of interest to note the flexible roles of $\tau^2_{i,0}$ and $\tau^2_{i,1}$, $\tau^2_{i,0}$ (resp. $\tau^2_{i,1}$) is freely estimated from the data using the absent (resp. present) prior connections. We allow $\tau^2_{i,0} < \tau^2_{i,1}$ which accommodates (rare) situations in which a prior is complementary to the data.

3. Variational Bayes method and Gibbs sampling In this section we develop a variational Bayes approach to approximate the (marginal) posterior distributions of the parameters $\beta_{i,r}, \tau^2_{i,0}, \tau^2_{i,1}, \sigma^2_i$ in model (4). The algorithm is similar, but still significantly different, from the algorithm developed in Leday et al. (2015) for the model (3). In the following we can see that, due to (4), the variational parameters have a form which renders the implementation of (4) much more challenging. We also verify that these approximations are accurate by comparing them to the results obtained using a Gibbs sampling strategy, which is much slower. Computational efficiency is an important characteristic, especially for fitting large networks.

In this section we work on a single regression equation, i.e. for a fixed index i, and given hyperparameters a_k, b_k, for $k = 0, 1, 2$. In the next section we combine the regression equations to estimate the hyperparameters.

3.1. Variational Bayes inference. In general a "variational approximation" to a distribution is simply the closest element in a given target set $\mathcal{Q}$ of distributions, usually with "distance" measured by Kullback-Leibler divergence. The set $\mathcal{Q}$ is chosen both for its computational tractability and accuracy of approximation. Distributions Q with stochastically independent marginals (i.e. product laws) are popular, and then the "accuracy" of approximation is naturally restricted to the marginal distributions.

In our situation we wish to approximate the posterior distribution of the parameter $\theta := (\beta_i, \tau^2_{i,0}, \tau^2_{i,1}, \sigma^2_i)$ given the prior (4) and the observation Y_i given in (2), for a fixed i. Here in (2) we take X_i (which depends on Y_j for $j \neq i$) as given, as in a fixed-effects linear regression model. For $p(\cdot\,|\,Y_i)$ the posterior density in this model, the variational Bayes approximation is given as

$$q^* = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathbf{E}_q \log \frac{q(\theta)}{p(\theta\,|\,Y_i)},$$

where the expectation is taken with respect to the density $q \in \mathcal{Q}$. For $p(Y_i, \theta)$ the joint density of (Y_i, θ), this is equivalent to finding the maximizer of

$$\mathbf{E}_q \log \frac{p(Y_i, \theta)}{q(\theta)}. \tag{5}$$

By the nonnegativity of the Kullback-Leibler divergence, the latter expression is a lower bound on the marginal density $p(Y_i) = \int p(Y_i, \theta)\, d\theta$ of the observation, and it is usually referred to as "the lower bound". Solving the variational problem is equivalent to maximizing this lower bound (over $\mathcal{Q}$).

We choose the collection $\mathcal{Q}$ equal to the set of distributions of θ for which the components β_i, $\tau^2_{i,0}$, $\tau^2_{i,1}$ and σ^2_i are stochastically independent, i.e. $q(\theta) = \prod_{l=1}^4 q_l(\theta_l)$, where the marginal densities q_l are arbitrary. Given such a factorization of q it can be shown in general (see e.g. Ormerod and Wand (2010)), that the optimal marginal densities q_l^* satisfy:

$$q_l^*(\theta_l) \propto \exp(\mathbf{E}_{q_{\setminus l}} \log p(Y_i, \theta)), \qquad \text{where } \mathbf{E}_{q_{\setminus l}} = \mathbf{E}_{q_1} \ldots \mathbf{E}_{q_{l-1}} \mathbf{E}_{q_{l+1}} \ldots \mathbf{E}_{q_4}.$$

It can be shown (see the Supplementary Material) that in model (4) for regression equation i, with $\theta = (\beta_i, \tau^2_{i,0}, \tau^2_{i,1}, \sigma_i^{-2})$, this identity can be written in the "conjugate" closed-form

$$\begin{aligned} \beta_i|\, Y_i &\sim \mathrm{N}\big(\beta_i^*, \Sigma_i^*\big), \\ \tau^2_{i,0}|\, Y_i &\sim \mathrm{Gamma}\big(a^*_{i,0}, b^*_{i,0}\big), \\ \tau^2_{i,1}|\, Y_i &\sim \mathrm{Gamma}\big(a^*_{i,1}, b^*_{i,1}\big), \\ \sigma_i^{-2}|\, Y_i &\sim \mathrm{Gamma}\big(a^*_{i,2}, b^*_{i,2}\big), \end{aligned} \tag{6}$$

where

$$\begin{aligned} \Sigma_i^* &= \big[\mathbf{E}_{q_4^*}(\sigma_i^{-2})\big(X_i^T X_i + \mathbf{D}_{\mathbf{E}_{q_2^*\cdot q_3^*}(\tau_i^2)}\big)\big]^{-1}, \\ \beta_i^* &= \big[X_i^T X_i + \mathbf{D}_{\mathbf{E}_{q_2^*\cdot q_3^*}(\tau_i^2)}\big]^{-1} X_i^T Y_i, \end{aligned}$$

$$\begin{aligned} a^*_{i,0} &= a_0 + \tfrac{1}{2}s^0, & b^*_{i,0} &= b_0 + \tfrac{1}{2}\mathbf{E}_{q_4^*}(\sigma_i^{-2})\mathbf{E}_{q_1^*}({\beta_i^0}^T\beta_i^0), \\ a^*_{i,1} &= a_1 + \tfrac{1}{2}s^1, & b^*_{i,1} &= b_1 + \tfrac{1}{2}\mathbf{E}_{q_4^*}(\sigma_i^{-2})\mathbf{E}_{q_1^*}({\beta_i^1}^T\beta_i^1), \\ a^*_{i,2} &= a_2 + \tfrac{1}{2}n + \tfrac{1}{2}s, & b^*_{i,2} &= b_2 + \tfrac{1}{2}\mathbf{E}_{q^*_{\setminus 4}}\big(\beta_i^T\mathbf{D}_{\tau_i^2}\beta_i\big) \\ & & & \quad + \tfrac{1}{2}\mathbf{E}_{q_1^*}(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i), \end{aligned}$$

where s^0 and s^1 are the number of 0's and 1's in the i-th row of the adjacency matrix P, not counting the diagonal element; and $\beta_i^0 = \{\beta_{i,r} : r \in \mathcal{I}\backslash i, \mathrm{P}_{i,r} = 0\}$ and $\beta_i^1 = \{\beta_{i,r} : r \in \mathcal{I}\backslash i, \mathrm{P}_{i,r} = 1\}$ are the coordinates of the vector of regression parameters corresponding to these 0's and 1's. Furthermore

$$\mathbf{D}_{\mathbf{E}_{q_2^*\cdot q_3^*}(\tau_i^2)} = \mathrm{diag}\Big(\mathbf{E}_{q_2^*}\mathbf{E}_{q_3^*}(\tau^2_{i,1}), ..., \mathbf{E}_{q_2^*}\mathbf{E}_{q_3^*}(\tau^2_{i,s})\Big).$$

In these identities the optimal densities q_l^* appear both on the left and the right of the equations and hence the identities describe the optimal densities only as a fixed point. In practice the identities are iterated "until convergence" from suitable starting values.

The iterations also depend on the hyperparameters a_k, b_k. In the next section we describe how these parameters can be estimated from the data by incorporating updates of these parameters in the iterations.

3.2. Variational Bayes vs Gibbs sampling. Under the true posterior distribution the coordinates $\beta_i, \tau^2_{i,0}, \tau^2_{i,1}, \sigma_i^2$ are not independent. This raises the question how close the variational approximation is to the true posterior distribution. As the latter is not available in closed form, we investigate this question in this section by comparing the variational approximation to the distribution obtained by running a Gibbs sampling algorithm. As for the network reconstruction we only use the marginal posterior distributions of the regression parameters, we restrict ourselves to these marginal distributions.

The full conditional densities of BSEMed can be seen to take the explicit form:

$$\begin{aligned} \beta_i|\, Y_i, \tau^2_{i,0}, \tau^2_{i,1}, \sigma_i^{-2} &\sim \mathrm{N}(\beta_i^*, \Sigma_i^*), \\ \tau^2_{i,0}|\, Y_i, \beta_i, \tau^2_{i,1}, \sigma_i^{-2} &\sim \mathrm{Gamma}(a^*_{i,0}, b^*_{i,0}), \\ \tau^2_{i,1}|\, Y_i, \beta_i, \tau^2_{i,0}, \sigma_i^{-2} &\sim \mathrm{Gamma}(a^*_{i,1}, b^*_{i,1}), \\ \sigma_i^{-2}|\, Y_i, \beta_i, \tau^2_{i,0}, \tau^2_{i,1} &\sim \mathrm{Gamma}(a^*_{i,2}, b^*_{i,2}), \end{aligned}$$

where the parameters Σ_i^*, β_i^*, $a^*_{i,k}$ and $b^*_{i,k}$ satisfy the same system of equations as in the variational algorithm, except that all expectations $\mathbf{E}_{q^*}$ must be replaced by the "current" values taken from the conditioning (see Supplementary Material). Thus Gibbs sampling of the full posterior $(\beta_i, \tau^2_{i,0}, \tau^2_{i,1}, \sigma_i^{-2})|\, Y_i$ is easy to implement, although slow.

We ran a simulation study with a single regression equation (say $i = 1$) with $n = p = 50$, and compared the variational Bayes estimates of the marginal densities with the corresponding Gibbs sampling-based estimates. Thus we sampled $n = 50$ independent replicates from a $p = 50$-dimensional normal distribution with mean zero and $(p \times p)$-precision matrix Ω, and formed the vector Y_1 and matrix X_1 as indicated in (2). The precision matrix was chosen to be a *band matrix* with a lower bandwidth b_l equal to the upper bandwith b_u. It is $b_l = b_u = 4$, thus a total number of 9 band elements including the diagonal. For both the variational approximation and the Gibbs sampler we used prior hyperparameters $a_2 = b_2 = 0.001$ and prior hyperparameters $\hat{a}_0, \hat{b}_0, \hat{a}_1, \hat{b}_1$ fixed to the values set by the *global* empirical Bayes method described in Section 4. The Gibbs iterations were run $nIter = 100,000$ times, after which the first $nBurnin = 1000$ iterates were discarded. Histograms based on subsampling every 10th value of the iterations are compared with the variational Bayes approximation to the marginal posterior densities. The correspondence between the two methods is remarkably good (see the Supplementary Material).

We conclude that the variational Bayes method gives reliable estimates of the posterior marginal distributions. The computing times in seconds are 40 for BSEMed and $2542 \times 50 = 127,100$ for the Gibbs sampling (in R). The variational method clearly outperforms the Gibbs sampling method, which would hardly be feasible even for $n = p = 50$.

4. Global empirical Bayes for BSEMed Model (4) possesses three pairs of hyperparameters (a_k, b_k), for $k \in \{0, 1, 2\}$. The pair (a_2, b_2) controls the prior of the error variances σ_i^2; we fix this to numerical values that render a vague prior, e.g. to $(0.001, 0.001)$. In contrast, we let the values of the parameters $\alpha = (a_0, b_0, a_1, b_1)$ be determined by the data. As these hyperparameters are the same in every regression model i, this allows information to be borrowed across the regression equations, leading to *global shrinkage* of the regression parameters.

A natural method to estimate the parameter α is to apply maximum likelihood to the marginal likelihood of the observations in the Bayesian BSEMed model determined by (2) and (4). Here "marginal" means that all parameters except α are integrated out of the likelihood according to their prior. The approach is similar to the one in van de Wiel et al. (2012). As a first simplification of this procedure we treat the vectors $Y_1, \ldots, Y_p$ as independent, thus leading to a likelihood of product form. As the exact marginal likelihoods of the Y_i are intractable, we make a second simplication and replace these likelihoods by the lower bound (5) to the variational Bayes criterion (see Supplementary Material).

Recall that in model (4) each regression parameter $\beta_{i,r}$ corresponds to one of two normal priors, that is:

$$\beta_{i,r} \sim \begin{cases} \mathrm{N}(0, \sigma_i^2 \tau_{i,0}^{-2}), & \text{if} \quad \mathrm{P}_{i,r} = 0, \\ \mathrm{N}(0, \sigma_i^2 \tau_{i,1}^{-2}), & \text{if} \quad \mathrm{P}_{i,r} = 1. \end{cases}$$

It is the regression coefficients corresponding to edges that are not present according to the prior information share the same precision $\tau_{i,0}^2$, and similarly the coefficients corresponding to the edges that are present obtain the precision $\tau_{i,1}^2$. Both precisions are assumed to have gamma priors with different hyperparameters that are adapted by the current data by the means of the global EB procedure described above. Then, if the absence of an edge in the prior network is supported by the current data, the corresponding regression coefficient $\beta_{i,r}$ will have a small variance, and will tend to be small; a similar, but opposite, situation will occur for edges that are present in the prior network. In next Section we shall see that the EB approach will tend to find similar values of $\tau_{i,0}^2$ and $\tau_{i,1}^2$ when the prior knowledge is non-informative, and rather different values otherwise.

We developed a dedicated edge selection algorithm for BSEM model in Leday et al. (2015). It is based on summarizing $\beta_{i,r}$ and $\beta_{r,i}$ by $\bar{\kappa}_{i,r}$,

$$\bar{\kappa}_{i,r} = (\kappa_{i,r} + \kappa_{r,i})/2 \quad \text{with} \quad \kappa_{i,r} = \frac{\left|\mathbf{E}_{q^{i*}}\left[\beta_{i,r}|\mathbf{y}_i\right]\right|}{\sqrt{\mathbf{V}_{q^{i*}}\left[\beta_{i,r}|\mathbf{y}_i\right]}} \tag{7}$$

where $\mathbf{E}_{q^{i*}}[\beta_{i,r}|\mathbf{y}_i]$ and $\mathbf{V}_{q^{i*}}[\beta_{i,r}|\mathbf{y}_i]$ denote the approximate posterior expectation and variance of $\beta_{i,r}$ obtained in Section 3. The $\bar{\kappa}_{i,r}$ values are ranked and corresponding edges are consecutively included according to a local false discovery rate (lfdr) criterion, which explores the relationship between lfdr and Bayes factors. Details are given in the Supplementary material.

5. Numerical investigation To study the effect of including a prior network in the model framework we compare BSEMed with BSEM. Hereto, we generated data $Y^1, \ldots, Y^n$ according to (1), for $p = 100$ and $n \in \{50, 200\}$, which reflect a high- and a low-dimensional situation, respectively. We considered precision matrices Ω_p, which imply *band*, *cluster* and *hub* network topologies (Zhao et al., 2012) (See Supplementary Material).

For BSEMed we vary the quality of the prior network information: 'perfect' prior information, i.e. the generating model; '75%' true edges; '50%' true edges; '0%' true edges. To generate 75% (or 50%, or 0%) true information, we swapped 25% (or 50%, or 100%) of the true edges with the same number of absent edges, i.e. in the adjacency matrix P that describes the prior network we swapped these percentages of 1s with 0s. It may be noted that in the last case the prior network is completely wrong for the true edges, but not for the absent edges due to over-sampling of the 0's, which seems realistic. Each simulation is repeated 50 times. We display the performances of BSEM and BSEMed by ROC curves, as based on ranking $\bar{\kappa}_{i,r}$, which summarizes $\beta_{i,r}$ and $\beta_{r,i}$ (7) (see Figure 1). We observe from Figure 1 that BSEMed performs better than BSEM when the prior information agrees the data and as good as BSEM when the prior doesn't. The latter reflects the adaptive nature of the EB procedure.

We also consider the EB estimates. We summarize the precisions by their prior means, as estimated by the EB procedure: $E(\tau^2_{i,k}) = \hat{a}_k/\hat{b}_k$, for $k \in \{0, 1\}$. When there is some agreement of the prior knowledge with the data, we expect $\hat{a}_0/\hat{b}_0 > \hat{a}_1/\hat{b}_1$. In the case with 0% true edges, the prior is partly wrong: none of the truly present edges are in the prior network while some of the truly absent edges are part of the prior network. Hence, we expect the EB procedure to produce $\hat{a}_1/\hat{b}_1$ that are slightly larger than $\hat{a}_0/\hat{b}_0$. As discussed in Section 2 for the complementary case, reversal of the roles of the two priors can still improve performance of BSEMed, or at least not deteriorate it.

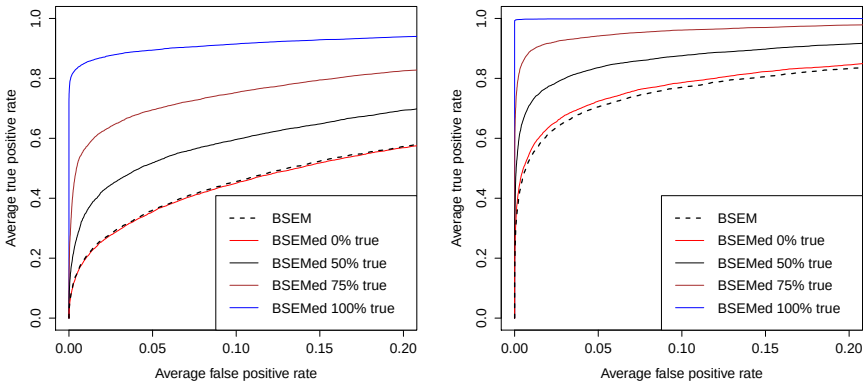
The EB estimates of the prior means are presented in Table 1 for the case corresponding to Figure 1(a): *band* structure, $n = 50$.

	$\hat{a}_0/\hat{b}_0$	$\hat{a}_1/\hat{b}_1$	ratio
true	366.10	8.08	45.30
0.75% true edges	272.97	14,36	19.00
0.50% true edges	216.10	27.56	7.84
0% true edges	142.59	152.95	1.07

Table 1 EB estimates of the prior means of precisions $\tau^2_{i,0}$ and $\tau^2_{i,1}$ in case of the *band* structure and $n = 50$ for various qualities of prior information

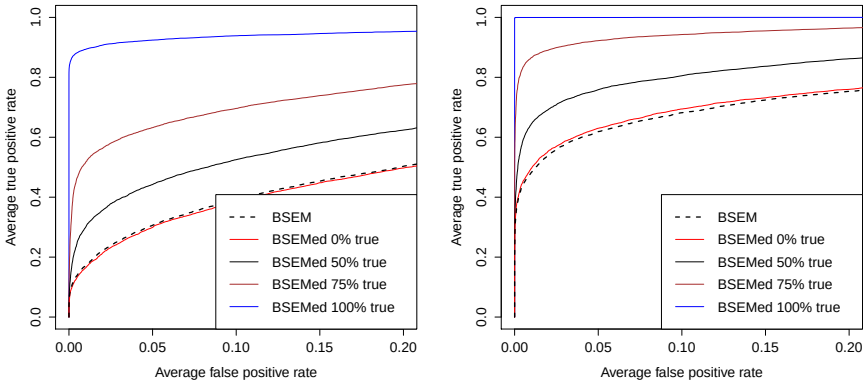
Table 1 displays the prior means of precision, as estimated by EB, for BSEMed models with various qualities of prior information. It is clear that the better the quality of the prior information is, the larger the ratio of mean prior precisions is. Tables for other simulation settings are available in the Supplementary material. These generally show the same pattern.

Figure 2 displays BSEM and BSEMed estimates of $\beta_{i,r}$ (3) and (4) for the *band* structure when $n = 50$ and $p = 100$ using the R package *rags2ridges* (Peeters and van Wieringen, 2014; van Wieringen and Peeters, 2014). Figures 1 & 2 show that BSEMed estimates become more accurate when prior knowledge quality increases and are as good as BSEM estimates when using 0% true edges information. It is also easy to see (Figure 2) a convergence of the BSEMed estimates to the true graph when the prior knowledge quality increases.



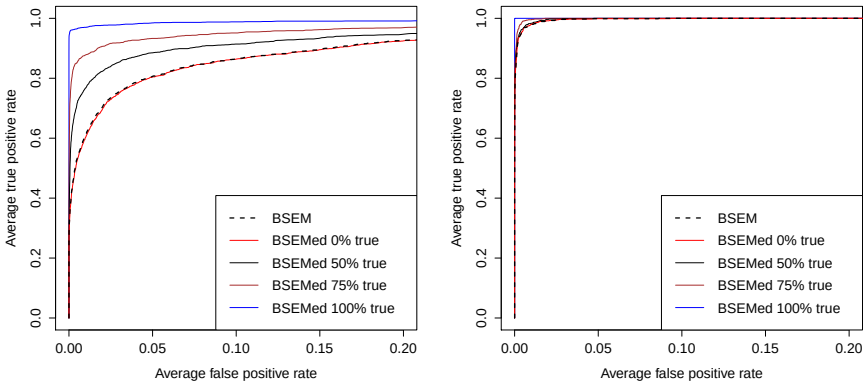
(a) Band: n = 50

(b) Band: n = 200



(c) Cluster: n =50

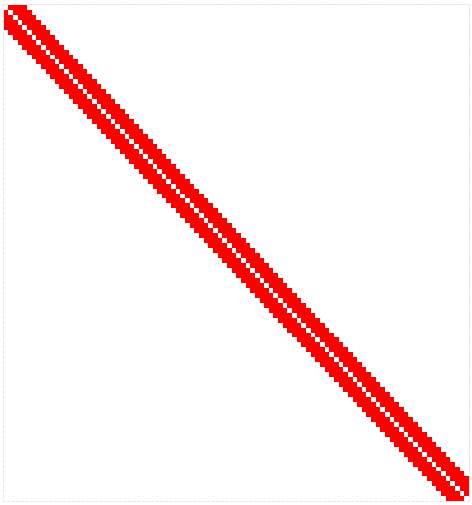
(d) Cluster: n = 200



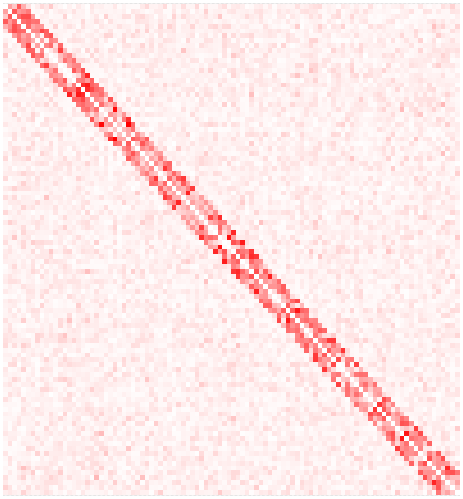
(e) Hub: n = 50

(f) Hub: n = 200

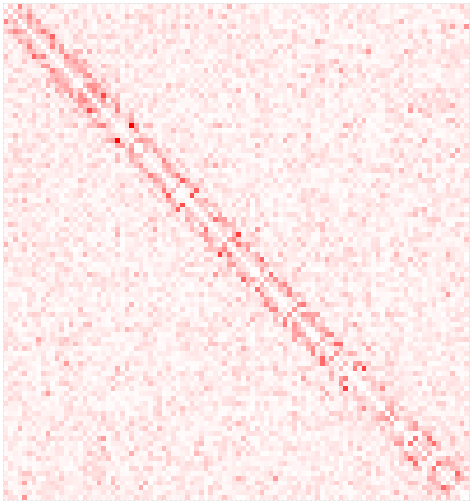
Figure 1 ROC curves for BSEM (dashed) and BSEMed using perfect prior information (blue), BSEMed using 75% true edges present in the prior (brown), BSEMed using 50% true edges present in the prior (black) and BSEMed using 0% true edges present in the prior (red). Here, $p = 100$ and $n \in \{50, 200\}$.



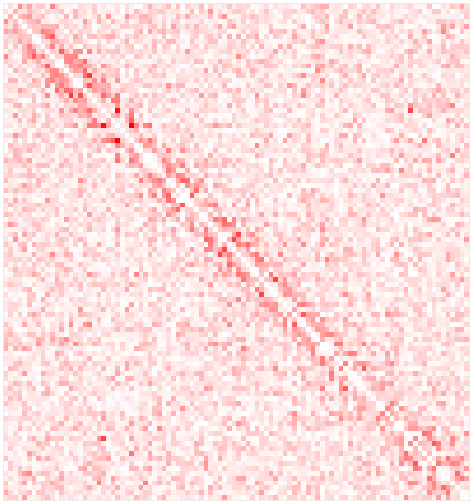
(a) True graph



(b) BSEMed: perfect prior



(c) BSEMed: 50 % true Info



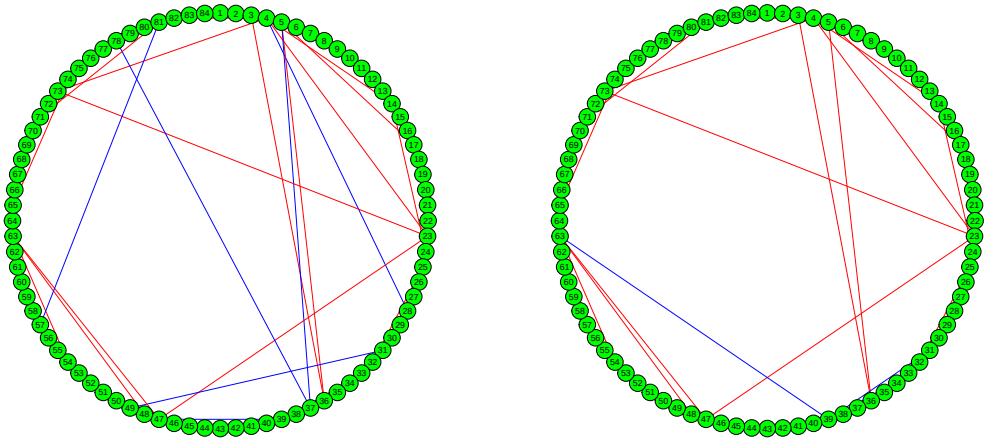
(d) BSEM

Figure 2 Visualization of BSEMed '$\bar{\kappa}_{i,r}$' using perfect prior (b), BSEMed '$\bar{\kappa}_{i,r}$' using 50% true edges information (c), BSEM '$\bar{\kappa}_{i,r}$' (d) and the true graph (a) in case $n = 50$ and $p = 100$.

6. Illustration We turn to real data in this section. We use gene expression data from the Gene Expression Omnibus (GEO) to illustrate and evaluate methods for reconstructing gene networks. We consider two types of cancer and cancer-related pathways. First, we focus on the Apoptosis pathway with $p = 84$ genes in a lung data set (Landi et al., 2008), consisting of $n_1^{\text{lung}} = 49$ observations from normal tissue and $n_2^{\text{lung}} = 58$ observations from tumor tissue, so $n^{\text{lung}} = 107$ in total. Secondly, we considered the p53 pathway in a pancreas data set (Badea et al., 2008) with $p = 68$ genes, consisting of $n_1^{\text{pancreas}} = 39$ observations from normal tissue and $n_2^{\text{pancreas}} = 39$ observations from tumor tissue, hence $n^{\text{pancreas}} = 78$ in total. Note that the data were scaled per gene prior to the computations.

BSEMed, BSEM, Graphical Lasso (GL_λ) (Friedman et al., 2008), SEM with the Lasso penalty (SEM_L) (Meinshausen and Bühlmann, 2006) and GeneNet (Schäfer et al., 2006) were applied on the tumor data parts of the data sets. For BSEMed, the corresponding data parts from normal tissue were used as prior knowledge by fitting genes networks on the normal data using BSEM. The idea is that, while tumors and normal tissue may differ quite strongly in terms of mean gene expression, the gene-gene interaction network may be relatively more stable.

We first illustrate the results from BSEM and BSEMed. Before considering the edge selection, we compare the total log-marginal likehood, as estimated by the variational lower bound, across the regression models for BSEM (3) and BSEMed (4) as a measure for goodness-of-fit. For the lung data set (resp. pancreas data set) we obtained -7082.93 for BSEM and -7071.99 for BSEMed (resp. -3807.58 for BSEM and -3798.91 for BSEMed). These improvements are clearly larger than what may be expected under random prior information of the same size, as shown in Supplementary Material in Section 7.

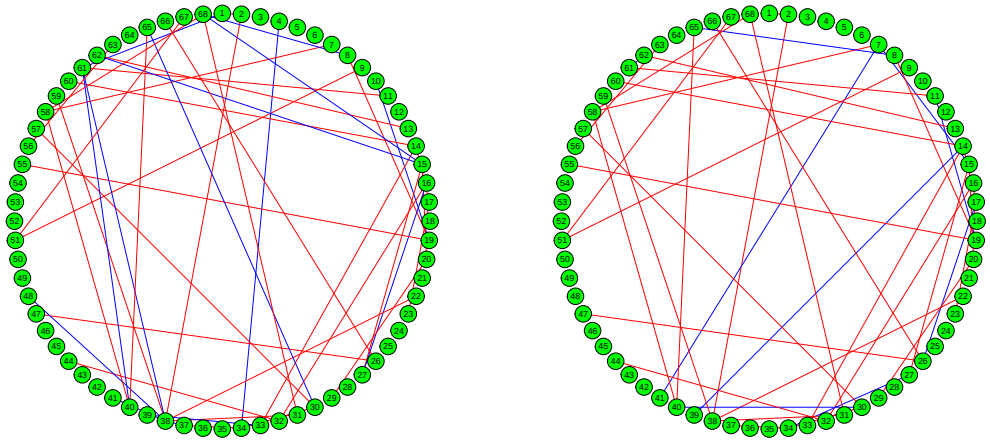


(a) BSEM network estimate (b) BSEMed network estimate

Figure 3 BSEM vs BSEMed network estimates in lung cancer. Red edges are the overlap edges.

Figure 3 (Figure 4) displays the estimated gene-gene network interaction in lung cancer (pancreas cancer) and their overlaps using the described selection procedure with estimated lfdr ≤ 0.1. Considerable overlap (red edges), but also notable differences can be seen.

Table 2 displays the prior means of precision, as estimated by EB. The prior network is clearly of use: the mean prior precision for regression parameters corresponding to the edges absent in the prior network is relatively large, which effectuates stronger shrinkage towards zero than for parameters corresponding to edges present in the prior network.



(a) BSEM network estimate (b) BSEMed network estimate

Figure 4 BSEM vs BSEMed network estimates in pancreas cancer. Red edges are the overlap edges.

	$\hat{a}_0/\hat{b}_0$	$\hat{a}_1/\hat{b}_1$	ratio
Lung	27.32	1.71	15.97
Pancreas	20.03	1.21	12.97

Table 2 EB estimates of precisions $\tau^2_{i,0}$ and $\tau^2_{i,1}$ of prior distributions in lung data (resp. pancreas data) set.

In the following, we argue that BSEMed network estimates may be more reliable in this setting than those of BSEM, Graphical Lasso (GL_λ) (Friedman et al., 2008), SEM with the Lasso penalty (SEM_L) (Meinshausen and Bühlmann, 2006) and GeneNet (Schäfer et al., 2006) (see the Supplementary Material for methodological details). For that, we assess performance of all methods by studying reproducibility of edges. We randomly split the tumor data part of the lung data set (pancreas data set) into two equal and independent parts: $n^{\text{lung}}_{2,1}$ and $n^{\text{lung}}_{2,2}$ (resp. $n^{\text{pancreas}}_{2,1}$ and $n^{\text{pancreas}}_{2,2}$). BSEM, BSEMed, GL_λ, GeneNet and SEM_L were applied on each subset of the tumor data. We repeated the procedure 50 times. We report in Table 3 (Table 4) the average number of overlapping edges between the two subsets for each method when the total number of edges selected by each method on each subset is set to 50, 100 and 200.

# edges	BSEM overlap	GeneNet overlap	SEM_L overlap	GL_λ overlap	BSEMed overlap	# prior edges in BSEMed
50	4.56	1.88	1.32	3.42	29.58	13.4
100	10.68	5.7	5.64	7.86	37.88	22.14
200	24.16	17.2	16.46	18.14	51.54	33.7

Table 3 Lung data, reproducibility study: Average number of overlapping edges among the top 50 (100, 200) strongest ones in two equally-sized splits of the tumor data for BSEMed, BSEM, GL_λ, GeneNet and SEM_L.

# edges	BSEM overlap	GeneNet overlap	SEM_L overlap	GL_λ overlap	BSEMed overlap	# prior edges in BSEMed
50	7.42	3.32	2.8	4.52	27.82	11.92
100	17.46	10.34	9.08	11.4	57.18	29.22
200	44.14	30.94	28.54	33.66	81.66	54.1

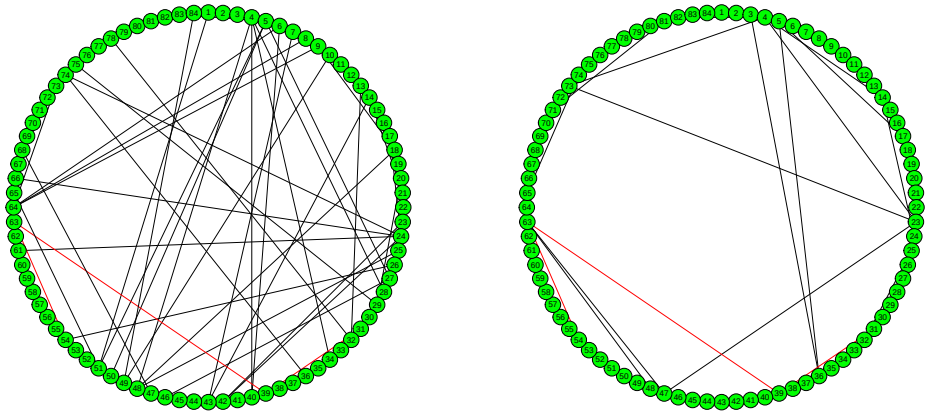
Table 4 Pancreas data, reproducibility study: Average number of overlapping edges among the top 50 (100, 200) strongest ones in two equally-sized splits of the tumor data for BSEMed, BSEM, GL_λ, GeneNet and SEM_L.

We observe from Tables 3 & 4 that the results from the BSEMed networks are much more reproducible than that of BSEM, which is on its turn more reproducible than the other ones. Clearly, the improvement can partly be explained by overlapping edges that were also part of the prior network. However, it is clear from Figure 5 that the BSEMed network estimate in tumor tissue is not just a 'finger print' of the prior network (normal tissue network): BSEMed can even reveal edges that are neither in prior network nor in BSEM network estimate.

(a) Lung data (b) Pancreas data

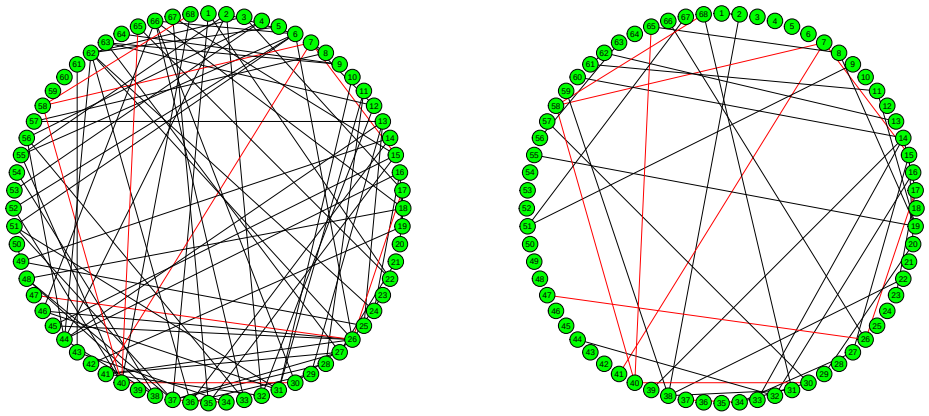
Figure 5 Venn diagrams displaying the mean overlap of reproduced top-ranking edges, corresponding to the second row of Table 3 (Figure 5.a) and Table 4 (Figure 5.b).

Figure 6 (resp. Figure 7) displays the network in normal tissue against the network in tumor tissue in the lung data (resp. in the pancreas data). The purpose of displaying Figure 6 and 7 is to emphasize the dysregulation of gene-gene interactions in cancer (Vogelstein and Kinzler, 2004; van Wieringen and van der Vaart, 2015) which may be caused by the heterogeneity of cancer (Nowell, 1976). Heterogeneity of tumor samples makes it more difficult to pinpoint reliable links, hence our selection algorithm which is based on local fdr ≤ 0.1 is likely to select fewer links in cancer samples.



(a) Network estimate in Normal tissue (b) BSEMed network estimate in tumor tissue

Figure 6 Network in a normal cell vs BSEMed network in lung cancer. Red edges are the overlap edges between prior and posterior networks.



(a) Network estimate in Normal tissue (b) BSEMed network estimate in tumor tissue

Figure 7 Network in a normal cell vs BSEMed network in pancreas cancer. Red edges are the overlap edges between prior and posterior networks.

7. Discussion We have presented a new method for incorporating prior information in undirected network reconstrustion based on Bayesian SEM. Our approach allows the use of two central Gaussian distributions per regression equation for coefficients $\beta_{i,r}$'s of our SEMs, where the prior information determines which of the two applies to a specific $\beta_{i,r}$. Empirical Bayes estimation of the parameters of the two hyper priors of the precisions introduces shrinkage and accommodates the situation where there would

not be an agreement of the prior information with the data at hand. We showed in simulation with different graph structures that BSEMed outperforms BSEM when the used prior knowledge (partially) agrees with the data and as good as when not. In addition, for two real data sets we showed better reproducibility of top ranking edges with respect to other methods .

In some cases, it may be desirable to give more weight only to some important edges of the prior graph rather than the whole graph. In gene regulatory networks reconstruction particularly, this may be edges that are known to characterise the disease biology. Assuming one is able to express such prior information as prior probabilities on edges, our software is able to incorporate such information via the Bayes factors used in the post-hoc selection procedure (Leday et al., 2015). Likewise, a user can also increase the weight of the entire prior graph uniformly (See Supplementary Material for details).

Instead of assigning Gaussian distributions to the coefficients, other (e.g. sparse) priors can be used. However, the fast variational Bayes method for posterior density approximation may not be valid anymore. For instance, would one use Horseshoe priors (Carvalho et al., 2010), the variational marginals are non-existent. The complement property (Section 2) is preserved whenever the same functional forms of the priors are used for both classes. However, a combination of e.g. a Gaussian and a sparse prior ruins this property, which renders such a combination less attractive.

Future research also focuses on extending our method to situations with more than two classes. For example, when considering integrative networks for two sets of molecular markers or two (related) pathways, the three class setting is relevant: two classes represent the connections within the two sets and a third one between the two sets. Finally, multiple sources of external data may be available for incorporation in BSEMed. This requires to model the parameter(s) of the priors in terms of contibutions of those external sources, and weigh those sources in a data-driven manner, as it is unlikely that the sources are equally informative.

Acknowledgements The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

Conflict of Interest

The authors have declared no conflict of interest.

References

Allen, G. I. and Liu, Z. (2013). A Local Poisson Graphical Model for Inferring Networks From Sequencing Data. *NanoBioscience, IEEE Transactions on* **12**, 189–198.

Badea, L., Herlea, V., Dima, S. O., Dumitrascu, T. and Popescu, I. (2008). Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology* **55**, 2016–2027.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

Chen, S., Witten, D. M. and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102**, 47–64.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks *In Proceedings of the IEEE Computer Science Bioinformatics Conference (CSB'03)*, IEEE, Washington DC, USA, pp. 104–113.

Isci, S., Dogan, H., Ozturk, C., and Otu, H. H. (2013). Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics* , 860–867.

Landi, M. T., Dracheva, T., Rotunno, M., Figueroa, J. D., Dasgupta A., Liu, H., Mann, F. E., Fukuoka, J., Hames, M., Bergen, A. W., Murphy, S. E., Yang, P., Pesatori, A. C., Consonni, D., Bertazzi, P. A., Wacholder, S., Shih, J.

H., Caporaso, N. E., and Jen, J. J. (2008). Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS ONE* **3**, e1651.

Lauritzen, S. (1996). *Graphical Models*. The Clarendon Press, Oxford University Press, New York .

Leday, G. G. R., de Gunst, M., Kpogbezan, G. B., van der Vaart, A. W., van Wieringen, W. N., and van de Wiel, M. A. (2015). Gene network reconstruction using global-local shrinkage priors. *arXiv preprint arXiv:1510.03771[stat.ME]*. To appear in *The Annals of Applied Statistics*.

Li, S., Wu, L., and Zhang, Z. (2006). Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach *Bioinformatics* **22**, pp. 2143—2150.

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34**, 1436–1462.

Mohammadi, A. and Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal.* **10**, 109–138.

Mukherjee, S., and Speed, T. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences* **105**, 14313–14318.

Nowel, P. C. (1976). The clonal evolution of tumor cell populations. *Sciences* **194**, 23–28

Omerod, J., and Wand, M. (2010). Explaining variational approximations. *The American Statistician* **64**, 140–153.

Peeters, C. F. W., and van Wieringen, W. N. (2014). *rags2ridges: Ridge estimation of precision matrices from high-dimensional data*. R package version 1.4

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using l1-regularized logistic regression. *Ann. Statist.* **38**, 1287—1319.

Schäfer, J., Opgen-Rhein, R., and Strimmer, K. (2006). Reverse engineering genetic networks using the GeneNet package. *R News* **6**, 50–53.

Steele, E., Tucker, A., 't Hoen, P.A.C., and Schuemie, M. J. (2009). Literature-based priors for gene regulatory networks *Bioinformatics* **25**, pp. 1768—1774.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

van de Wiel, M. A., Leday, G. G. R., Pardo, L., Rue, H., van der Vaart, A. W., and van Wieringen, W. N. (2012). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* .

van Wieringen, W. N. and Peeters, C. F. W. (2014). Ridge estimation of inverse covariance matrices from high-dimensional data. *arXiv preprint arXiv:1403.0904[stat.ME]*.

van Wieringen, W. N. and van der Vaart, A. W. (2015). Transcriptomic heterogeneity in cancer as a consequence of dysregulation of the genegene interaction network. *Bull. Math. Biol.* **77**, 1768—1786 .

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine* **10**, 789–799.

Werhli, A. and Husmeier, D. (2007). Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology* **6**.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

Yang, E., Ravikumar, P., Allen, G. and Liu, Z. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems 25* (P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 1367—1375.

Zhao, T., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2012). The **huge** package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13**, 1059—1062.

